Online Research @ Cardiff

This is an Open Access document downloaded from ORCA, Cardiff University's institutional repository: https://orca.cardiff.ac.uk/id/eprint/125791/

This is the author's version of a work that was submitted to / accepted for publication.

Citation for final published version:

Dongyu, She, Jufeng, Yang, Ming-Ming, Cheng, Lai, Yukun ORCID: https://orcid.org/0000-0002-2094-5680, Rosin, Paul ORCID: https://orcid.org/0000-0002-4965-3884 and Liang, Wang 2020. WSCNet: Weakly Supervised Coupled Networks for Visual Sentiment Classification and Detection. IEEE Transactions on Multimedia 22 (5) , pp. 1358-1371. 10.1109/TMM.2019.2939744 file

Publishers page: http://dx.doi.org/10.1109/TMM.2019.2939744 <http://dx.doi.org/10.1109/TMM.2019.2939744>

Please note:
Changes made as a result of publishing processes such as copy-editing, formatting and page numbers may not be reflected in this version. For the definitive version of this publication, please refer to the published source. You are advised to consult the publisher's version if you wish to cite this paper.

This version is being made available in accordance with publisher policies. See http://orca.cf.ac.uk/policies.html for usage policies. Copyright and moral rights for publications made available in ORCA are retained by the copyright holders.

information services
gwasanaethau gwybodaeth

WSCNet: Weakly Supervised Coupled Networks for Visual Sentiment Classification and Detection

Dongyu She, Jufeng Yang, Ming-Ming Cheng, Yu-Kun Lai, Paul L. Rosin and Liang Wang

***Abstract*—Automatic assessment of sentiment from visual content has gained considerable attention with the increasing tendency of expressing opinions online. In this paper, we solve the problem of visual sentiment analysis, which is challenging due to the high-level abstraction in the recognition process. Existing methods based on convolutional neural networks learn sentiment representations from the holistic image, despite the fact that different image regions can have different influence on the evoked sentiment. In this paper, we introduce a weakly supervised coupled convolutional network (WSCNet). Our method is dedicated to automatically selecting relevant soft proposals given weak annotations (*e.g.*, global image labels), thereby significantly reducing the annotation burden, and encompasses the following contributions. First, the proposed WSCNet detects a sentiment-specific soft map by training a fully convolutional network with the cross spatial pooling strategy in the detection branch. Second, both the holistic and localized information are utilized by coupling the sentiment map with deep features as semantic vector in the classification branch. The sentiment detection and classification branches are integrated into a unified deep framework optimized in an end-to-end manner. Extensive experiments demonstrate that the proposed WSCNet outperforms the state-of-the-art results on seven benchmark datasets.**

***Index Terms*—Visual sentiment analysis, weakly supervised detection, convolutional neural networks**

I. Introduction

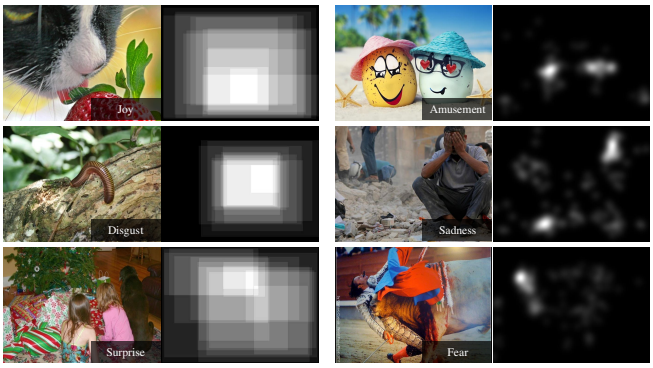
Visual sentiment analysis from images has attracted great attention with an increasing tendency of expressing opinions via posting images on social media platforms, *e.g.*, Flickr and Twitter. Assigning image sentiment automatically has various applications, *e.g.*, affective computing [2], opinion mining [3], [4], emotion-based image retrieval (EBIR) [5], [6], entertainment [7], [8], *etc*. Recently, due to the success of convolutional neural networks (CNNs), numerous deep approaches have been proposed to predict sentiment [9], [10]. The effectiveness of machine learning based deep features has been demonstrated over hand-crafted features (*e.g.*, color, texture, and composition) [11]–[13] on visual sentiment prediction. However, several issues exist when using CNNs to address such an abstract task, which are explained as follows.

D. She, J. Yang and M.-M. Cheng are with the College of Computer Science, Nankai University, Tianjin 300350, China (e-mail: sherry6656@163.com; yangjufeng@nankai.edu.cn; cmm@nankai.edu.cn).

Y.-K. Lai and P.L. Rosin are with the School of Computer Science and Informatics, Cardiff University, Wales, UK (e-mail: laiy4@cardiff.ac.uk; Paul.Rosin@cs.cf.ac.uk).

L. Wang is with the National Laboratory of Pattern Recognition, CAS Center for Excellence in Brain Science and Intelligence Technology, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: wangliang@nlpr.ia.ac.cn).

A preliminary version of this work appeared in CVPR [1].



(a) (b)

Fig. 1. Examples from the (a) EmotionROI [14] and (b) EMOd datasets [15] with the human annotation. The ground-truth sentiments are also given. The normalized maps indicate the regions that influence the evoked sentiments and emotional attention. As can be seen, the sentiments can be evoked by specific regions.

First, compared with conventional recognition tasks, visual sentiment analysis is more challenging due to a higher level of subjectivity in the human recognition process [2]. Neuroimaging and behavioral studies find that human attention is attracted by emotional relevance of a stimulus [16]–[18], which is also proved as the *emotion prioritization effect* in computer vision studies [15]. Fig. 1 shows examples from the EmotionROI [14] and EMOd datasets [15]. As can be seen, specific regions show strong influence on evoked sentiment. It is necessary to take such an effect into consideration for visual sentiment prediction, while most existing methods employ CNNs to learn representations only from entire images [19], [20]. Second, precise annotations (*e.g.*, bounding boxes) can provide more discriminative information than image-level labeling, which also lead to better performance in recognition tasks [21]. However, there are two limitations for visual sentiment classification using region-based annotations. On the one hand, collecting such precise annotations can be very labor-intensive and time-consuming, whereas achieving only image-level annotations is much easier, especially for such a subjective task. On the other hand, different regions contribute differently to the viewer's evoked sentiment, while crisp proposal boxes only tend to find the foreground objects in an image.

To address these problems, this paper proposes a weakly supervised coupled network (WSCNet) framework for joint sentiment detection and classification with two branches, namely detection and classification branches. The first branch

is designed to generate region proposals evoking sentiment. Instead of extracting multiple crisp proposal boxes, we use a soft sentiment map to represent the probability of evoking the sentiment for each receptive field. In detail, we make use of a fully convolutional network (FCN) followed by the proposed cross-spatial pooling strategy to summarize the feature maps into image-level scores. Thus, the network can be trained with image-level sentiment labels, which significantly reduces the annotation burden. Then the sentiment map is generated and utilized to highlight the regions of interest that are informative for classification. In addition, the second branch captures the localized representation by coupling the sentiment map with the deep features, which is then combined with the holistic representation to provide a more semantic vector.

Our contributions are summarized as follows: First, we introduce a weakly supervised coupled network integrating visual sentiment classification and weakly supervised sentiment detection into a unified CNN framework, which learns the discriminative representation for visual sentiment analysis in an end-to-end manner. Second, we exploit sentiment maps to provide image-specific localized information with only image-level labels, with which both holistic and localized representations are fused as semantic vector for robust sentiment classification. Extensive experiments demonstrate that the proposed framework performs favorably against the state-of-the-art methods on seven benchmark datasets.

This paper is an extended version of our conference paper [1], to which we enrich the contributions in the following four aspects: (1) We provide useful details of our weakly supervised framework, and distinguish it from comparative methods, *e.g.*, salience detection and weakly supervised detection frameworks. (2) We add a comprehensive review of related work making the manuscript more self-contained. (3) We conduct an exhaustive analysis on the weakly supervised detection framework for visual sentiment prediction and add evaluation on the eye-tracking dataset [15]. For comparison, recent learning based salience detection methods are also trained and evaluated on such datasets. (4) We carefully study the capability and failure mode of our approach, and highlight the difference between the sentiment map and other attention and salience work.

TABLE I
STATISTICS OF THE AVAILABLE AFFECTIVE DATASETS. MOST DATASETS DEVELOPED IN THIS FIELD CONTAIN A FEW THOUSAND SAMPLES, MAINLY DUE TO THE SUBJECTIVE AND LABOR INTENSIVE LABELING PROCESS. AS THE LAST COLUMN SHOWS, NONE OF THESE DATASETS EXCEPT EMOTIONROI AND EMOD PROVIDE GROUND TRUTH REGIONS THAT EVOKE SENTIMENTS.

Dataset	#Images	#Classes	Regions
IAPSa [12]	395	8	N
Abstract [12]	228	8	N
ArtPhoto [12]	806	8	N
Twitter I [19]	1,269	2	N
Twitter II [31]	603	2	N
EmotionROI [14]	1,980	6	Y
EMOd [15]	1,019	10	Y
Flickr&Instagram [10]	23,308	8	N
Flickr [43]	60,745	2	N
Instagram [43]	42,856	2	N

II. RELATED WORK

Our work is closely related to two recent trends in computer vision community, *i.e.*, understanding and recognition of visual sentiment, and weakly-supervised detection algorithms.

A. Visual Sentiment Prediction

The literature on visual emotion prediction can be roughly divided into categorical and dimensional approaches. The categorical approaches [2], [22], [23] identify sentiments with a limited set of categories according to psychological studies, *e.g.*, sadness, fear. Likewise, with a dimensional view of sentiments, the dimensional approaches place sentiment in a two- or three-dimensional space, *e.g.*, valance-arousal (VA) [24], [25], which allow for a greater range of expressions. This paper mainly focuses on categorical approaches aiming at mapping sentiments into intuitive categories [26]–[28]. In the early years, there are numerous methods using hand-crafted features for image sentiment classification [11], [29], [30]. For example, Machajdik *et al.* [12] define a combination of rich hand-crafted features based on art and psychology theory, *e.g.*, composition, color variance and image semantics, while Zhao *et al.* [13] introduce more robust and invariant visual features designed according to art principles. Moreover, Borth *et al.* [31] propose the visual sentiment ontology (VSO) and detectors to detect adjective noun pairs (ANP) from images as a mid-level representation, while a similar method in [29] leverages the mid-level attributes. To predict personalized emotion perceptions, Zhao *et al.* [32], [33] further propose the multi-task hypergraph learning, considering different factors that may influence emotion perceptions, *i.e.*, visual content, social context, temporal evolution and location influence. However, such hand-crafted visual features are shown to be effective only on several small datasets [34], while having limitations in classifying large-scale images from social media.

More recently, CNN-based approaches [10], [35]–[37] have also been applied to recognize visual sentiments and achieve significant advances. For example, Chen *et al.* [38] construct DeepSentiBank as a visual concept of ANP classification model. Campos *et al.* [20], [39] fine-tune state-of-the-art CNNs pre-trained on the large-scale general dataset [40] for visual sentiment prediction. You *et al.* [41] and Wu *et al.* [41] propose to make use of web images for training deep models due to lack of well-labeled data. In addition, there are several methods utilizing the rich information from the multiple layers of CNN models. Zhu *et al.* [35] propose a unified CNN-RNN framework to integrate different levels of features by exploring their dependencies, while Rao *et al.* [42] propose a multi-level deep network to unify both low-level and high-level information in images.

Psychology study findings indicate that human attention usually prioritizes emotional content (*e.g.*, smiling babies) over emotionally neutral stimuli [16], [46], [47]. While most CNN-based methods for sentiment classification extract deep features from the entire image, significantly less attention has been paid to utilize the localized information for sentiment

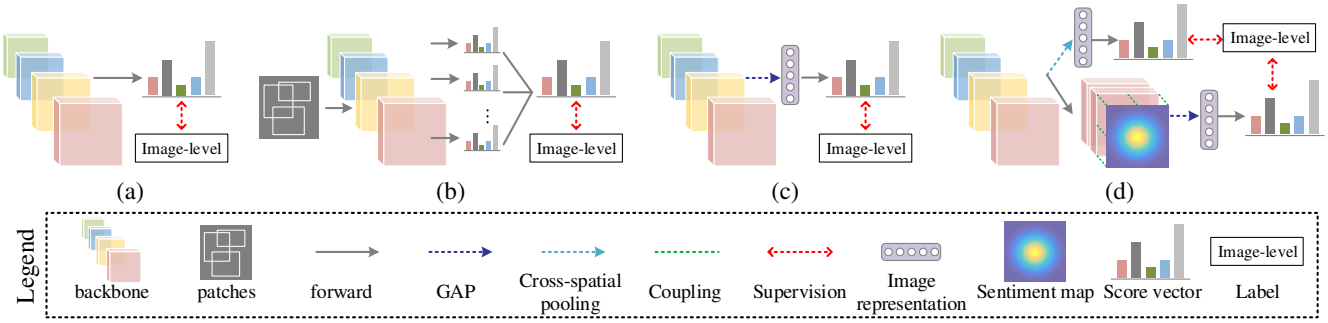


Fig. 2. Illustration of different architectures. (a) plain CNN, (b) MIL [44] (c) CAM [45], (d) our proposed architecture. In (a), a fixed-size image is fed into the CNN. In (b), a set of patches only requires image-level annotations for training, while the candidate bounding boxes are generated in multiple stages. In (c), the object localization information comes from viewing the convolutional filters as detectors in a unified network, which is ignored for learning sentiment representation. The proposed WSCNet in (d) introduces cross-spatial pooling for summarizing the information from the deep feature maps and combines the advantages of utilizing both holistic representation and localized information.

prediction [48]. Recently, Sun *et al*. [49] and Yang *et al*. [50] discover affective regions based on an off-the-shelf object proposal algorithm and combine deep features for classification. However, such methods are sub-optimal since the objectness algorithm is separate from the prediction method, and regions that are not object-like may be excluded at the very beginning. In [51], a method based on an attention model is developed in which local visual regions induced by sentiment related visual attributes are considered. In addition, Peng *et al*. [14] train a supervised network FCNEL to predict the emotion stimuli map (ESM) with manually labeled pixel-level ground truth. Fan *et al*. [52] investigate how attention influences visual sentiment and further propose a novel DNN model with a subnetwork that is able to encode the relative importance of regions within an image [15]. However, such fully supervised methods would be extremely labor intensive if they were extended to large-scale datasets. The existing datasets in this field are summarized in Tab. I, most of which only contain limited samples with image-level annotation. Different from existing methods in the literature, we propose a weakly supervised model to learn a discriminative sentiment representation for both classification and detection. Experimental results show the superiority of the proposed framework over the state-of-the-art methods.

B. Weakly Supervised Detection

With the recent success of deep learning on large-scale object recognition [53], several weakly supervised CNNs have been proposed for the object localization task [54], [55]. The objective of these methods is to localize object parts that are visually consistent with the semantic image-level labels across the training data. One of the most common approaches for tackling this task is to formulate it as a multiple instance learning (MIL) problem [44], [56]–[59]. MIL defines images as a bag of regions, and assumes that images labeled as positive contain at least one object instance of a certain category and images labeled as negative do not contain an object from the category of interest, as shown in Fig. 2 (b). Cinbis *et al*. [60] consist of generating object proposals and extracting features from the proposals in multiple stages, and employ MIL on the features to determine the box labels from the weak bag labels. In [61], a weakly-supervised deep learning pipeline is proposed to localize objects from complex cluttered scenes by explicitly searching over possible object locations and scales in the image. Since the training process of the MIL alternates the stages of object extraction and classifier training, the solutions are non-convex and as a result are sensitive to the initialization.

Recently, some studies show a similar intuition that CNNs trained using weak supervisions can provide object location information, which try to localize objects by first generating object score heatmaps and then placing bounding boxes around the high response regions [45], [55], [62], [63]. For example, Zhou *et al*. [45] address weakly-supervised object localization using global average pooling and extend their analysis to abstract concepts, which provides a typical solution in this domain. As shown in Fig. 2 (c), they aggregate class-specific activation maps (CAM) by adding a global average pooling (GAP) layer. Porzi *et al*. [64] introduce the top-N average pooling to find the best compromise between average and max pooling for urban scenes, while Alameda-Pineda *et al*. [65] further propose LENA pooling layer for virality recognition. Similarly, Durand *et al*. [66] propose Wildcat pooling to summarize all information contained in the feature maps for each class, which takes maximum and minimum scoring regions into consideration, while Zhu *et al*. [62] also propose soft proposal networks (SPN) to generate soft proposals for weakly supervised object localization. In addition, Wei *et al*. [67], [68] propose to mine dense object regions for supervision by incorporating the adversarial erasing into finding the corresponding semantic regions. Zhang *et al*. [55] further propose to generate Self-produced Guidance (SPG) masks for improving the localization performance. However, such weakly supervised methods are mainly tested on the object localization dataset, which contains a large portion of natural iconic-object images, *i.e*., a single large object located in the center of an image or several different objects located separately. Considering that affective images contain ambiguous sentiment, where different emotions may

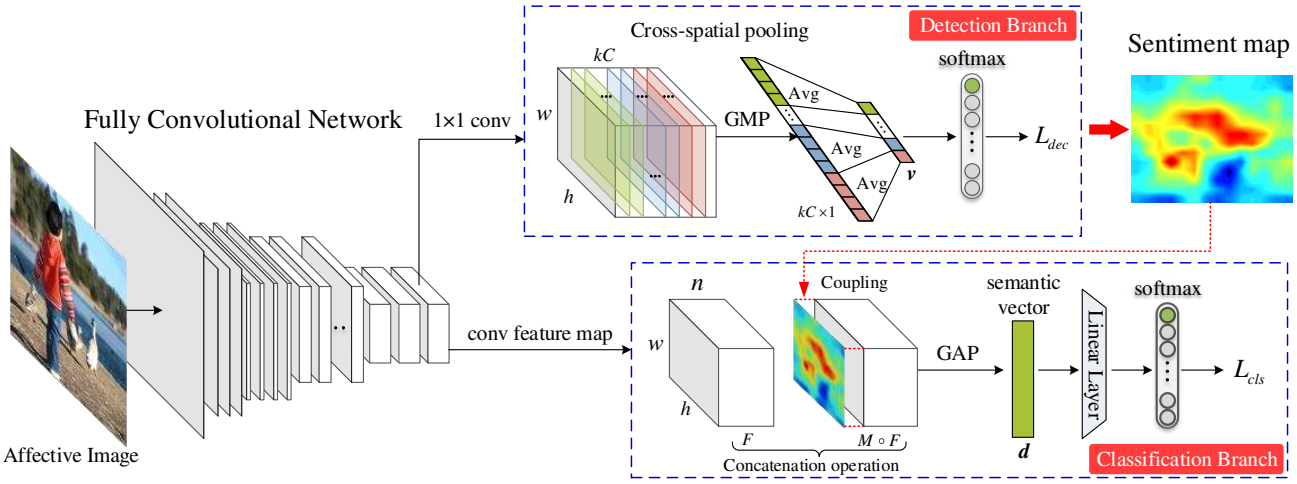


Fig. 3. Illustration of the proposed WSCNet for visual sentiment analysis. The input image is first fed into the convolutional layers of FCN, and the response feature maps are then delivered into two branches. The detection branch employs the cross-spatial pooling strategy to summarize all the information contained in the feature maps for each class. The generated sentiment map is then coupled with the original deep feature maps in the classification branch, resulting in the localized representation of the input image. Then, both holistic and localized representations are concatenated as a semantic vector for sentiment classification. These two branches only require image-level supervisions for training.

coexist in the same stimuli, the performance of these methods can be limited for such subjective tasks.

In this paper, we follow this research direction and analyze whether such intuition that views the convolutional filters as detectors to activate locations is still effective when used for classifying and localizing patterns associated with visual sentiment. Different from the existing methods, as shown in Fig. 2 (d), we integrate sentiment-related proposals into CNNs for utilizing local information under weak supervision. Instead of using class-specific activation [45], [66], this work detects a unified sentiment map considering all the activation maps by a weighted sum pooling strategy, due to the ambiguous information between the sentiments. Moreover, the detected sentiment map is coupled on the feature maps, which are then combined with the global representation as a more semantic vector. Thus, the detection and classification branches can boost each other during the end-to-end training process.

III. Weakly Supervised Coupled Network

Fig. 3 illustrates the proposed weakly supervised coupled network, which aims to detect soft proposals that evoke sentiment, only requiring image-level labels as the manual supervision. Specifically, WSCNet jointly optimizes both detection and classification tasks with two network branches, *i.e.*, detection branch and classification branch. The detection branch is employed to generate a sentiment map providing the localized information, which is then fed into the classification branch, fusing the holistic as well as the localized representations to form the semantic vector for classification.

A. Sentiment Map Detection Branch

A sentiment image is defined as a person's disposition to respond to visual inputs according to the psychological theory [69]. While attention and salience works aim to find salient objects in images, this paper focuses on the regions evoking sentiment, which may contain not only salient objects but other related areas [14]. As mentioned above, there are only a few end-to-end CNN frameworks for weakly supervised object detection that do not use additional localization information. In order to infer the sentiment map directly in the CNN, the convolutional filters are viewed as the detector that produces the feature maps as the response. Different from the object detection methods that employ the RoI pooling [70] operation on the bounding box [71]–[73], a form of soft proposal is used to represent the probability of evoking the sentiment for each receptive field. We first propose a cross-spatial pooling strategy to summarize the feature maps to the categorization-level information.

Cross-spatial pooling strategy. For a collection of N training examples $\{(x_i, y_i)\}_{i=1}^{N}$, let x_i denote an affective image, $y_i \in \{1, \cdots, C\}$ denotes the corresponding sentiment label, and C is the number of affective categories. For each instance, let $F \in \mathbb{R}^{w \times h \times n}$ be the feature maps of the last convolutional layer in the CNN, where w and h are the spatial size (width and height) of the feature maps, respectively, and n is the number of channels. We first add a 1×1 convolutional layer to capture multiple information for each sentiment category, which has a high response to certain discriminative regions. Suppose k detectors are applied to each sentiment class, we obtain feature maps F' with the dimension of $w \times h \times kC$. We propose to summarize all the information as a single image-level score for each of the sentiment classes independently, regardless of the input size, which is achieved by the cross-spatial pooling strategy:

$$v_c = \frac{1}{k} \sum_{i=1}^{k} \boldsymbol{G}_{max}(f_{c,i}), c \in \{1, \cdots, C\}, \quad (1)$$

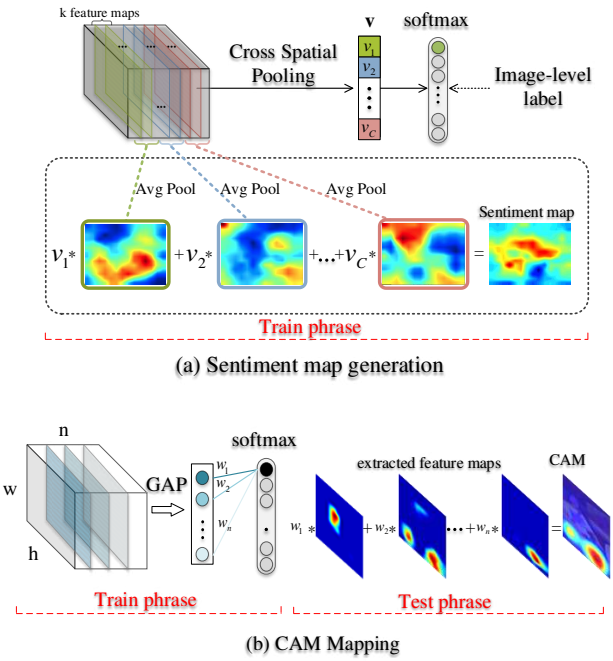


Fig. 4. Overview of the (a) sentiment map generation and (b) class activation mapping [45]. The sentiment map can be generated by mapping the predicted class scores of the input image to the deep feature maps during the training phrase, while the CAM needs an unnatural way for visualization only in the test phase by using the weights from the trained network.

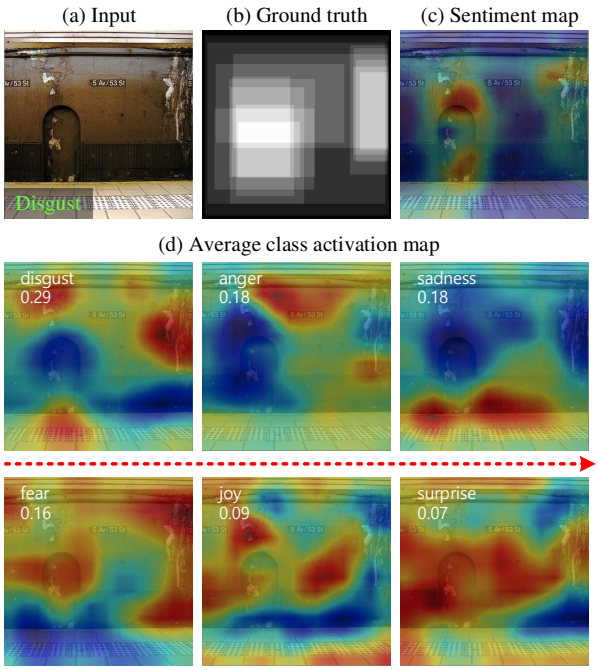


Fig. 5. The sentiment map generated from the top 6 classes for the given "disgust" image. The predicted class label and its score from the detection branch are shown in each activation map. We observe that the highlighted regions vary across predicted classes.

where $f_{c,i}$ represents the i-th feature map for the c-th label from F', and $\boldsymbol{G}_{max}(\cdot)$ denotes the Global Max Pooling (GMP). Here, GMP is employed to identify just one discriminative part for each feature map in the same sentiment class inspired by [45], which results in a $1 \times 1 \times kC$ vector. Then k responses for each label are unified with the average pooling operation, where the value can be maximized by finding all discriminative regions of the specific sentiment, as all low activations reduce the output of the particular map. The pooled vector $\mathbf{v} \in \mathbb{R}^C$ is then fed into a C-class softmax layer as the sentiment detection loss:

$$L_{dec} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C}\mathbf{1}(y_i = c)\log v_c, \tag{2}$$

where $\mathbf{1}(s) = 1$ if the condition s is true, and 0 otherwise. Thus, the filter weights can be updated during the training process, which yields the discriminative location in the feature maps for each class. We use the cross-spatial pooling strategy to represent the GMP layer followed by a class-specific average pooling as a convenient term.

Generating the sentiment map. Different from object locations [61] or class activation maps [45], the activation feature maps for different sentiments are dependent due to the ambiguity existing in the sentiment labels [36]. Thus, this paper proposes to capture the regions evoking sentiment by considering all the class activation maps with corresponding weights.

We first obtain a single map from the k feature maps for each sentiment, here the average pooling operation is employed to take multiple information into consideration. All the C class-wise feature maps with corresponding weights are then considered to capture the comprehensive localized information, instead of using the feature maps with the largest response from a specific class (see also Fig. 4 (a)). Thus, our sentiment map $M \in \mathbb{R}^{w \times h}$ is generated using v_c as the weight of the response map of class c:

$$M = \sum_{c=1}^{C} v_c \left(\frac{1}{k} \sum_{i=1}^{k} f_{c,i} \right). \tag{3}$$

Intuitively, based on prior methods [74], we expect that each unit v_c is activated by some visual patterns within its receptive field. The sentiment map is a weighted linear sum of the presence of these visual patterns at different spatial locations. By simply up-sampling the activation map to the size of the input image, we can identify the regions most relevant to the evoked sentiment.

B. Coupled Sentiment Classification Branch

The original convolutional feature can be viewed as the holistic representation from the perspective of image representation. While the sentiment map highlights the image-specific discriminative regions, such a map can be utilized to produce a local representation that is informative for image classification. Inspired by [62], the Hadamard product is employed to couple each feature map from the original feature

maps F with M. Thus, we obtain the coupled feature maps $U = [U_1, U_2, \cdots, U_n]$, where the element $U_i = M \circ F_i$, and $\circ$ denotes the element-wise multiplication. For fusing the multi-view information, we use vector fusion in the classification branch, which can benefit from end-to-end learning. Then the coupled feature maps and the original feature maps can be encoded to form a more informative semantic feature $\mathbf{d} \in \mathbb{R}^{2n}$ by:

$$\mathbf{d} = \boldsymbol{G}_{avg}(F \uplus U), \tag{4}$$

where $\uplus$ denotes the concatenation of different convolutional features. In the above equation, $\boldsymbol{G}_{avg}(\cdot)$ is the global average pooling (GAP) operation, which outputs the average value of each feature map.

We then add a fully-connected layer to compute the predicted scores of the input image for different classes. And the sentiment scores $s(y_i = c|\mathbf{d}, \mathbf{w}_c)$ are defined as:

$$s(y_i = c|\mathbf{d}, \mathbf{w}_c) = \frac{\exp(\mathbf{w}_c^\top \mathbf{d})}{\sum_{c'=1}^{C} \exp(\mathbf{w}_{c'}^\top \mathbf{d})}, \tag{5}$$

where $\mathbf{W} = \{\mathbf{w}_c\}_{c=1}^{C}$ is the set of model parameters. Thus, the classification is carried out by minimizing the following log likelihood function:

$$L_{cls} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} \mathbf{1}(y_i = c) \log s(y_i = c|\mathbf{d}, \mathbf{w}_c). \tag{6}$$

In this network, the C-way classification layer is determined by the number of classes in the affective dataset.

C. Joint Training Process

As shown in Fig. 3, our WSCNet will produce two outputs for sentiment detection and sentiment classification tasks. Given the training set, we explicitly train the proposed deep model to optimize the joint loss function:

$$L = L_{dec}(x, y) + L_{cls}(x, y). \tag{7}$$

Since derivatives w.r.t. all the parameters can be derived, we can conduct an effective end-to-end representation learned using stochastic gradient descent (SGD) to minimize the joint loss function. With this scheme, we can detect the sentiment map using weakly supervised learning, and utilize the localized information for discriminative classification.

D. Discussion

In order to utilize the image-level label for training, the cross-spatial pooling strategy is employed to summarize the information of feature maps into image-level scores, which includes no parameters to learn compared to others, *e.g.*, the attention-based strategy [77], [78]. This kind of architecture is also reversed in the CAM-based methods [45], [79], which employ global pooling before the last fully connected layer. For example, the whole network needs to be trained first, and fully-connected weights of the corresponding class are then extracted to combine the feature maps from the previous convolutional layer, as shown in Fig. 4 (b). This order needs an unnatural way for visualizing class-specific heatmaps, while the proposed cross-spatial pooling layer can be visualized with direct localization of discriminating regions. In addition, due to the ambiguity information existing in the sentiments, we generate the sentiment map taking all the response feature maps into consideration. In Fig. 5, we highlight the differences for utilizing different classes to generate the maps. Note that the sentiment scores reported are from the detection branch, corresponding to the pooled vector v_c. For the input disgust image, the high scores are all from related classes (*e.g.*, other negative sentiments like anger and sadness), providing the complementary information.

IV. Experiments

In this section, we evaluate the proposed WSCNet on visual sentiment classification and detection tasks. The datasets and experimental setup are described in Sec. IV-A and Sec. IV-B, respectively. We evaluate the effectiveness of our method for classification and discuss important parameters in Sec. IV-C. Finally, we evaluate the detection performance on two datasets and visualize the quality of detection results in Sec. IV-D and Sec. IV-E.

A. Datasets

We evaluate the proposed WSCNet on seven public affective datasets including the Flickr and Instagram (FI) [10], Flickr [43], Instagram [43], Twitter I [19], Twitter II [31], EmotionROI [14], EMOd datasets [15].

The FI dataset is labeled by a group of 225 Amazon Mechanical Turk (AMT) participants. Each one is asked to label the images from social websites that are queried with eight sentiment categories as keywords, *i.e.*, *anger, amusement, awe, contentment, disgust, excitement, fear, sadness*. And 23,308 images receiving at least three agreements finally form the dataset. The Flickr and Instagram datasets contain 60,745 and 42,856 images from Flickr and Instagram, respectively, each image of which is annotated with a sentiment polarity (*i.e.*, positive, negative) label. The above three datasets are the current largest datasets in the domain. In addition, we also evaluate on four small-scale datasets. Twitter I and Twitter II datasets are collected from the social websites and labeled with sentiment polarity categories by AMT participants, which consist of 1,269 and 603 images, respectively. The EmotionROI dataset is created for a sentiment prediction benchmark, which is assembled from Flickr resulting in 1,980 images with six sentiment categories. Besides, each image is also annotated with 15 regions that evoke sentiments, which are normalized to range between 0 and 1 as an emotion stimuli map (ESM) [14]. The EMOd dataset is constructed from two sources: (1) a subset (321) photos of the International Affective Picture System (IAPS); (2) a set of 698 photos collected by the authors. The EMOd dataset is the first to include eye-tracking data. Subject eye movements are recorded by asking sixteen subjects to observe each image freely for 3 seconds, followed by a drift correction that requires subjects to fixate at the screen center. For each image, a fixation map is generated by placing at each fixation location a Gaussian distribution with sigma equal to one degree of visual angle and then normalizing the map to have a maximum value of 1.

TABLE II
CLASSIFICATION ACCURACY (%) ON SEVEN AFFECTIVE DATASETS, INCLUDING FI, FLICKR, INSTAGRAM, TWITTER I, TWITTER II, EMOTIONROI, EMOD DATASETS. WE EVALUATE THE PROPOSED WSCNET AGAINST SEVERAL BASELINE METHODS INCLUDING THE EXTRACTED FEATURE BASED METHODS, DEEP LEARNING-BASED METHODS AND WEAKLY-SUPERVISED FRAMEWORKS. NOTE THAT SUN *et al*. AND YANG *et al*. METHODS ARE PROPOSED FOR BINARY CLASSIFICATION AND MULTI-CLASS CLASSIFICATION, RESPECTIVELY, AND THUS DATASETS WITH INCOMPATIBLE CLASS NUMBERS CANNOT BE EVALUATED, DENOTED AS '–'.

Method	FI	Flickr	Instagram	EmotionROI	Twitter I	Twitter II	EMOd
Zhao *et al*. [13]	46.13	66.61	64.17	34.84	67.92	67.51	17.20
SentiBank [31]	49.23	69.26	66.53	35.24	66.63	65.93	18.93
DeepSentiBank [38]	51.54	70.16	67.13	42.53	71.25	70.23	21.79
ImageNet-AlexNet [53]	38.26	69.05	56.69	34.26	65.80	67.88	20.68
ImageNet-VGG16 [75]	41.22	69.88	63.44	37.26	67.49	68.79	22.54
ImageNet-Res101 [76]	50.01	72.26	67.28	40.79	72.55	70.42	28.77
Fine-tuned AlexNet	58.13	73.11	69.95	41.41	73.24	75.66	40.13
Fine-tuned VGG16	63.75	78.14	77.41	45.46	76.75	76.99	43.21
Fine-tuned Res101	66.16	80.03	79.33	51.60	78.13	78.23	46.56
Sun *et al*. [49]	–	80.95	80.67	–	82.73	80.91	–
Yang *et al*. [36]	66.79	–	–	52.40	–	–	–
SPN [62]	66.57	79.71	79.53	52.70	81.67	77.96	47.25
WILDCAT [66]	67.03	80.67	80.31	55.05	79.53	78.81	46.83
CAM-Res101 [45]	68.54	79.21	79.46	55.72	82.67	79.13	46.08
Ours	**70.07**	**81.36**	**81.81**	**58.25**	**84.25**	**81.35**	**48.95**

TABLE III
CLASSIFICATION ACCURACY (%) OF WSCNET USING DIFFERENT NUMBERS OF FEATURE MAPS ON THE TEST SET OF THREE LARGE-SCALE DATASETS, *i.e*., FI, FLICKR, INSTAGRAM. IN THE REMAINING EXPERIMENTS, WE SET $k = 4$ IN OUR FRAMEWORK.

Dataset	$k = 1$	$k = 2$	$k = 4$	$k = 8$	$k = 16$
FI	68.23	69.36	70.07	68.80	67.19
Flickr	81.46	81.87	81.36	81.15	81.98
Instagram	79.67	79.24	81.81	79.60	78.53

TABLE IV
ABLATION STUDY ON THE FI DATASET. THE BASELINE IS WSCNET ($k = 1$) WITHOUT THE COUPLING OPERATION, DENOTED AS *Base*. NOTE THAT *SM* DENOTES USING THE SENTIMENT MAP AS THE GUIDANCE, *Local* DENOTES THAT ONLY THE COUPLED FEATURE MAP IS USED FOR CLASSIFICATION, AND *Coupling* DENOTES CAPTURING BOTH THE HOLISTIC AND LOCALIZED INFORMATION AS PRESENTED IN EQ. 4.

#	*Base*	$k = 4$	*SM*	*Local*	*Coupling*	FI
1	√					66.57
2	√	√				67.96
3	√		√	√		67.69
4	√		√		√	68.23
5	√	√	√		√	70.07

B. Experiment Setup

1) Implementation details: Our method is built on the pre-trained ResNet-101 [76] on the ImageNet dataset. To deal with the limited training data, we apply random horizontal flips and crop a random 448×448 patch as a form of data augmentation to reduce overfitting. We replace the last layers (global average pooling and fully connected layer) by the proposed multi-branch layer. The added layers are initialized using Gaussian distributions with mean 0 and standard deviations 0.01, and the biases are initialized to 0. The momentum and weight decay are set to 0.9 and 0.0005 respectively. During training, the mini-batch size for SGD is set to 32, the learning rates of the convolutional layers and the last fully-connected layers on both branches are initialized as 0.001, 0.01 respectively. The FI datasets are split randomly into 80% training, 5% validation and 15% testing sets. For the Flickr dataset and Instagram dataset, we randomly sample the same number of images for each class following the same configuration in [43], which are split randomly into 90% training, 10% testing sets. The small-scale datasets are split into 80% training and 20% testing sets randomly except those with specified training/testing splits [14], [31]. At test time we average the predictions of ten images (*i.e*., the five crops and their horizontal reflections) from the classification branch as final results. The sentiment map is extracted from the detection branch according to Eq. 3 as the probability of regions evoking sentiment for detection evaluation. Our framework is implemented based on the PyTorch deep learning framework [80]. All of our experiments are performed on an NVIDIA GTX Titan X GPU with 32 GB on-board memory.

2) Baseline: We evaluate the proposed WSCNet against thirteen baselines including methods using traditional features, CNN-based methods and weakly-supervised frameworks. For the traditional methods, we extract the principle-of-art features [13] from the affective images. We use a simplified version provided by the author to extract 27 dimensional features and use LIBSVM [84] for classification. We use the 1,200 dimensional mid-level representation from the ANP detector of SentiBank and apply the pre-trained DeepSentiBank to extract 2,089 dimensional features. For the basic CNN models, we report the results of using three classical deep learning methods pre-trained on ImageNet and fine-tuned on the affective datasets: AlexNet [53], VGGNet [75] with 16 layers and ResNet-101 [76]. We also show the results of fully-connected features extracted from the ImageNet CNN with LIBSVM. We use the default value and employ the *one v.s. all* strategy. We also report the results from three state-

TABLE V
EMOTIONAL ATTENTION PREDICTION (RANK) ON THE EMOD DATASET USING DIFFERENT METHODS, INCLUDING THE BASELINES, OBJECTNESS DETECTION ALGORITHM, SALIENCY DETECTION METHODS, WEAKLY SUPERVISED FRAMEWORKS AND THE SUPERVISED MODEL. NOTE '*' DENOTES THE CASNET IS PRE-TRAINED ON THE DATASET WITH FULL SUPERVISION. "AVG" INDICATES THE AVERAGE RANK OF EACH WEAKLY SUPERVISED METHOD. HERE, WE USE THE NORMALIZED FIXATION MAP AS THE GROUND-TRUTH FOR EVALUATION.

Algorithm	AUC-J	AUC-B	CC	SIM	KL	EMD	AVG
Random Map	0.50(7)	0.50(7)	0.00(7)	0.30(7)	1.88(3)	4.25(4)	5.83(6)
Center Crop	0.68(4)	0.59(4)	0.33(5)	0.39(4)	9.58(7)	3.71(3)	4.50(4)
Objectness [81]	0.61(5)	0.56(5)	0.17(6)	0.31(6)	7.51(6)	5.04(7)	5.83(6)
GBVS [82]	**0.80(1)**	**0.66(1)**	0.46(2)	0.47(2)	5.96(5)	4.59(6)	2.83(2)
IttiKoch [83]	0.73(3)	0.63(3)	0.37(3)	0.43(3)	2.09(4)	**3.16(1)**	2.83(2)
WILDCAT [66]	0.55(6)	0.52(6)	0.37(4)	0.32(5)	1.66(2)	4.52(5)	4.67(5)
WSCNet	0.76(2)	0.64(2)	**0.48(1)**	**0.48(1)**	**1.23(1)**	3.63(2)	**1.50(1)**
CASNet* [15]	0.86	0.72	0.64	0.56	0.85	1.98	-

of-the-art deep methods for sentiment classification. For the binary datasets, we use Sun's method [49] to select top-1 regions and combine the holistic feature with the region feature from the fine-tuned ResNet. For the multi-class datasets, we employ Yang's method [36] to transform the single label to a sentiment distribution and report the classification performance using ResNet. Moreover, we also evaluate our method against the state-of-the-art weakly supervised frameworks, *i.e.*, the WILDCAT, SPN, CAM methods, which are also based on ResNet-101 with the same input size of 448×448 as our method.

For the detection task, we evaluate the performance of sentiment map detection against different methods. Three baseline methods are employed to generate regions of interest for affective images, *i.e.*, random map, center crop and objectness region generated by [81] and faster RCNN [70]. To generate these baseline maps, we assign random probability to each pixel, crop half of the image from the center, and use the objectness tool [81] to generate one object region for each image. We also use the Graph-Based Visual Saliency model (GBVS) [82] and Itti-Koch model (IttiKoch) [83] to compute the saliency map. For the weakly supervised methods, we directly extract CAM (class activation maps) from the fine-tuned ResNet-101 following [45], and also evaluate against the final feature maps from the WILDCAT and SPN methods. In addition, two fully supervised methods, *i.e.*, FCNEL [14] and CASNet [15], are also tested on the EmotionROI and EMOd datasets, respectively, providing the upper bound for weakly supervised detection. Note that CASNet is trained on the SALICON [85] to achieve their best possible performance, directly tested on the EMOd without training/fine-tuning on them following [15].

3) Metrics: For the classification performance, we use the universally-agreed metric: accuracy. For evaluation of sentiment detection, we use four commonly used metrics: MAE, precision (P), recall (R), and $F-$score. Before the evaluation, we first binarize the predicted map using Otsu thresholding following [14]. MAE is the mean absolute error between the value of the predicted map and the ground truth map at all locations. The precision is defined as

$$P = \frac{1}{N}\sum_{i=1}^{N}\frac{|b_i \cap g_i|}{|b_i|}, \tag{8}$$

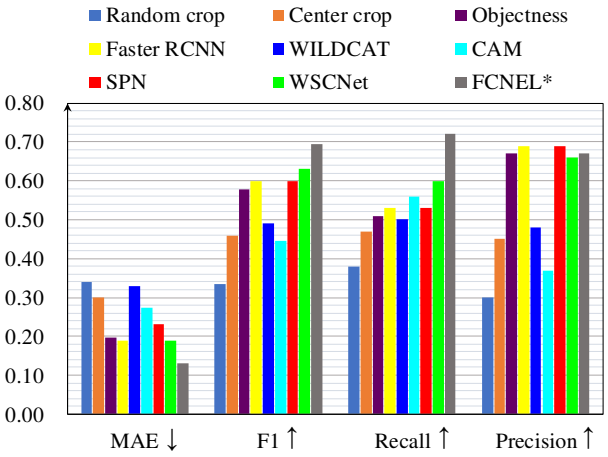


Fig. 6. Sentiment detection performance on the EmotionROI dataset (better viewed in color version). We compare our proposed WSCNet with algorithms including baseline methods, object detection algorithms, weakly supervised methods and the fully-supervised model FCNEL. Note that '*' denotes that the method is trained with bounding box annotation, providing the upper bound for weakly supervised detection methods. We employ four metrics for detection evaluation. For MAE, lower is better, denoted by ↓). For the others, larger is better, denoted by ↑.

where $|\cdot|$ is used to measure the number of pixels within the given set. Note that g_i and b_i are the ground truth emotion and the detected proposal of the i-th image. The recall is defined as

$$R = \frac{1}{N}\sum_{i=1}^{N}\frac{|b_i \cap g_i|}{|g_i|}. \tag{9}$$

Thus, $F-$score is computed using $F = 2 \times \frac{R \times P}{R+P}$. In addition, for the attention prediction, we use 6 metrics for comprehensive evaluation following [15], including two variants of AUC (Area Under the Curve) and four similarity metrics. AUC-J and AUC-B [86] treat the saliency map as a binary classifier, which alleviates the effects of center bias. Linear Correlation Coefficient (CC) [87], histogram intersection (SIM) [88], the Earth Movers Distance (EMD) [89] and the Kullback-Leibler divergence (KL) [90] are used to measure the similarity between the saliency map and fixation map. Note that for the

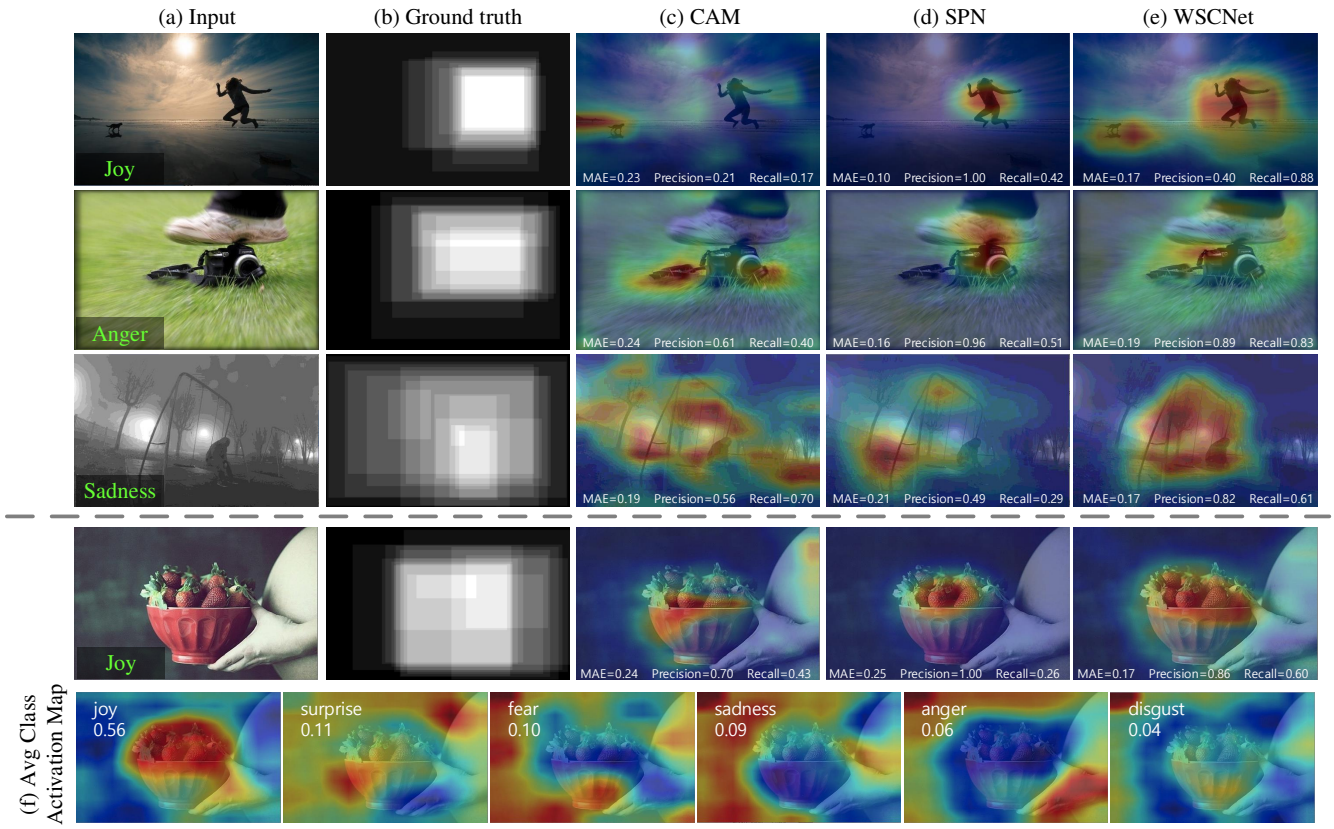


Fig. 7. Weakly supervised detection results using different methods on the EmotionROI dataset. The input images and the ground truth are given in (a) and (b). The detected regions and metric values of different weakly-supervised methods (*i.e.*, CAM, SPN, WSCNet) are shown in the last three columns. In the last row, we also show each average class activation maps, listed in descending order of predicted probability for each sentiment category. As can be seen, by activating the sentiment-related areas, our method achieves the most accurate results compared to the ground truth.

first four metrics, larger is better, while for the last two metrics smaller is better.

C. Classification Performance

We first evaluate the classification performance on seven affective datasets, followed by a detailed discussion. We set the hyper-parameter $k = 4$ in the proposed WSCNet as the default setting. Tab. II shows that the deep representations outperform the hand-crafted features, while the fine-tuned CNNs have the capability to recognize sentiment from images. The weakly supervised frameworks improve the performance of Fine-tuned Res101 utilizing the regional information. Our proposed method consistently performs favorably against the state-of-the-art methods for sentiment classification, *e.g.*, about 3.3% improvement on FI and 5.8% on EmotionROI, which illustrates that WSCNet can learn more discriminative representation for this task. The following are the detailed discussion for our proposed framework.

1) Hyper-parameter k: We first analyze the effect of the hyper-parameter k, *i.e.*, the number of the response feature maps for each sentiment category. Tab. III reports the classification performance of the detection branch using different k on the FI, Flickr, Instagram datasets. With an increasing number of feature maps, our method is able to achieve better performance compared with the standard classification strategy in the CNN (*i.e.*, $k = 1$), which captures multiple views for each sentiment category. However, over-amplifying the feature maps results in suboptimal performance mainly due to overfitting, which is similar to the finding reported in WILDCAT [66]. For the FI and Instagram datasets, our method achieves the best performance with $k = 4$, and for the Instagram dataset, the best performance is achieved with $k = 16$, although the performance is fairly stable with changing k. Therefore, we set $k = 4$ in our framework for a trade-off between efficiency and effectiveness.

2) Different Branch Accuracy: We report the classification performance of the classification and detection branches, since both branches use the image-level annotations for training. On the FI dataset, the classification branch achieves 70.07%, while the detection branch achieves a sub-optimal performance of 68.51%. When fusing features from the detection and classification branches, the LIBSVM result shows similar performance (70.18%) as the classification branch. Thus, we only use the classification branch as the final results.

3) Ablation Study: We perform an ablation study to illustrate the effect of each contribution. Our baseline is WSCNet with $k = 1$ and without the coupling operation, where the classification branch is the original classification layer in the CNN (*i.e.*, global pooling and fully connected layer). From Tab. IV, we can draw the following conclusions: First,

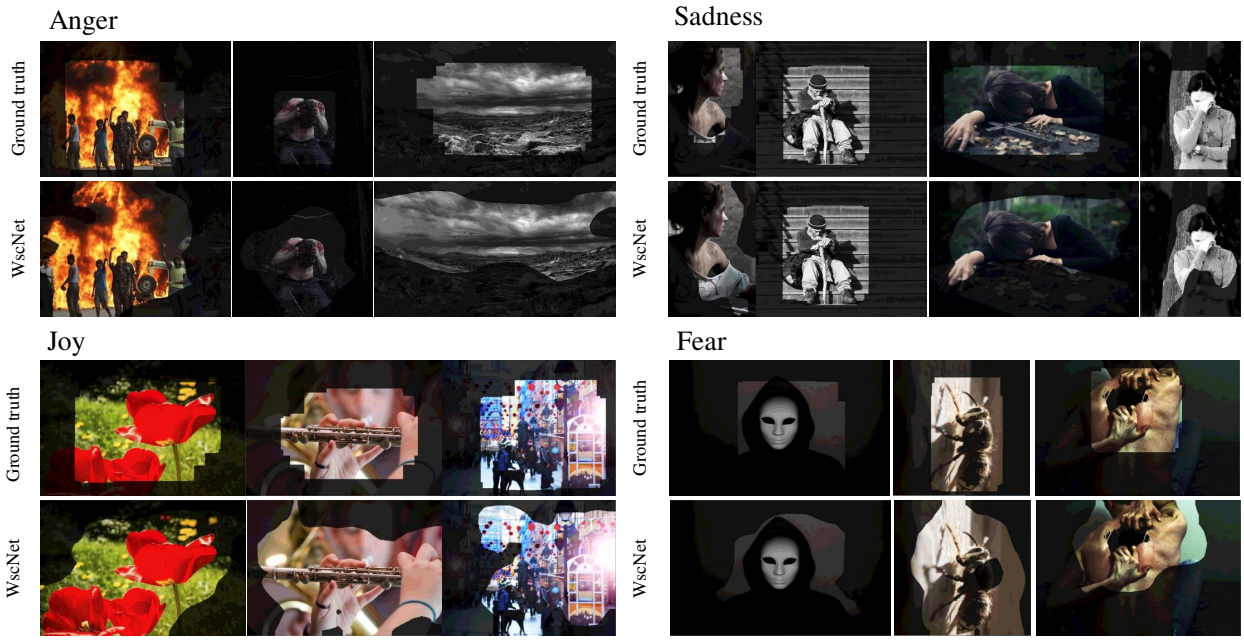


Fig. 8. Examples of class-specific units from the proposed WSCNet on the EmotionROI dataset [14]. Both binarized ground truths and sentiment maps are shown. The proposed weakly supervised method can achieve comparable results as the human annotations without the labeling burden.

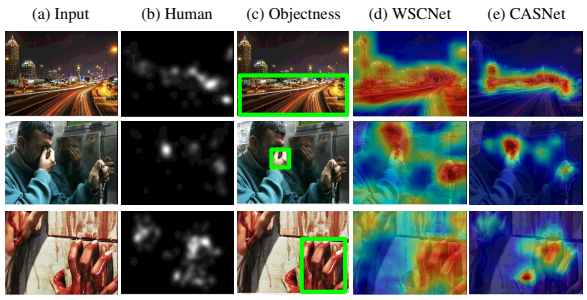


Fig. 9. Qualitative results generated by our model in comparison with the objectness methods and fully supervised method (CASNet) on the EMOd dataset.

using both multiple feature maps ($k = 4$) and the sentiment map coupled representation improve classification accuracy by about 1% on FI, illustrating the effectiveness of local representation. Second, utilizing the coupling operation combining multiple view information improves the base performance by 1.7%. Third, we achieve the best accuracy by utilizing the components to train our model in an end-to-end manner, which shows the complementarity of all the contributions.

D. Sentiment Detection

Fig. 6 reports the detection performance of different methods on the EmotionROI dataset. As shown, our WSCNet performs favorably against the baselines and weakly supervised methods (*i.e.*, WILDCAT, CAM, SPN), and also achieves comparable performance with the supervised FCNEL on most evaluation metrics. We notice that FCNEL benefits from supervised training with bounding box annotation, and has significantly better recall than other methods. The reason is that the regions evoking sentiments contain both the primary objects and additional contextual background, while Objectness [81] only focuses on the foreground objects and thus achieves a reasonable precision. Compared with the existing weakly supervised methods, our method improves the recall to 0.60, which illustrates the effectiveness of taking the sentiment characteristic into consideration for generating the sentiment map.

We also evaluate the performance on the EMOd dataset in Tab. V. We compare our method with six baselines without training on ground truth regions. As can be seen, the saliency models perform better on the AUC metrics, however, such metrics cannot distinguish between cases where models predict different relative importance values for different regions of an image [15]. The proposed WSCNet has the best overall performance among the baselines on the CC, SIM, KL metrics, as well as the average rank (AVG), demonstrating its advantage on emotional attention.

E. Visualization

We provide qualitative results in Figs. 7-10. We first show more detection results using different weakly supervised methods on the EmotionROI. As shown in Fig. 7, compared with the ground truth, WSCNet is able to detect the relevant regions that influence the evoked sentiment, while CAM and SPN may only focus on the salient objects leading to a reasonable precision but a low recall. For example, on the third row, SPN only responds to the foreground objects, which leads to

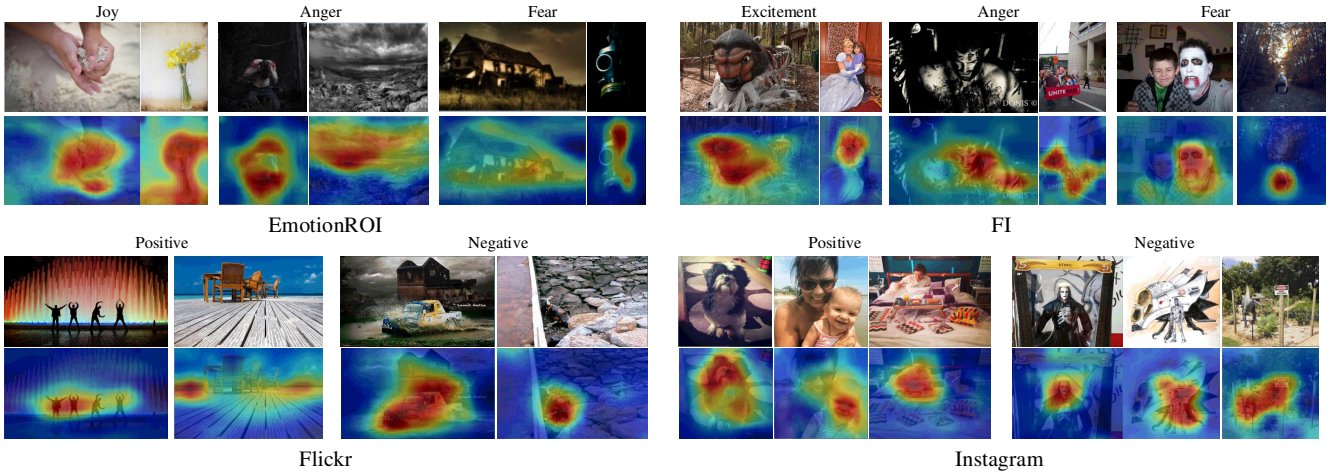


Fig. 10. Visualization of our detected sentiment maps on four datasets, *i.e.*, FI, EmotionROI, Flickr, Instagram. Our detected locations are not limited to foreground objects, but also include sentiment-related background.

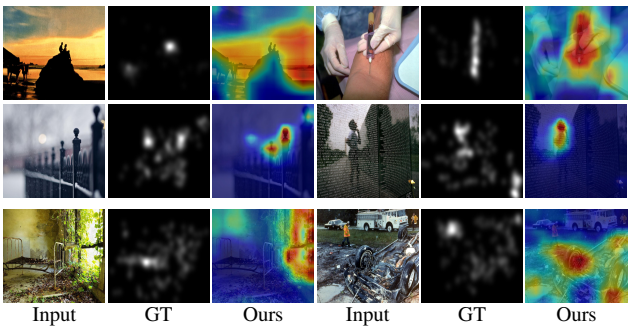


Fig. 11. Failure cases selected from the EMOd dataset. As can be seen, most cases are caused by small objects, low contrast between foreground and background, and complex background.

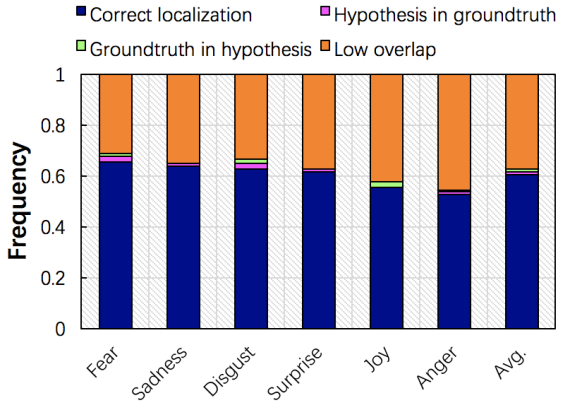


Fig. 12. Per-class frequency of error modes, averaged across all classes on the EmotionROI dataset.

0.96 precision but only 0.51 recall. In contrast, our detected sentiment map extends the object regions into the sentiment related background, which achieves the recall of 0.83. In addition, the last row shows more detailed results of the detected sentiment map. Specifically, we show each average class activation map as well as the corresponding weight, *i.e.*, the sentiment scores reported are from the detection branch, corresponding to the pooled vector v_c. As can be seen, although two maps of opposite sentiment classes focus on different regions of interest, by integrating the regions of interest from the related class, our proposed method is able to obtain a complementary sentiment map.

Fig. 8 shows the class-specific units for different sentiment categories on the EmotionROI dataset as in [45]. Both the detected sentiment maps as well as the ground truth are generated using the Otsu thresholding for binarization. From the figure we can identify the regions of the images that are most discriminative for classification and exactly which units detect these regions. The results show that the proposed weakly supervised method can achieve comparable results as the human annotations without the labeling burden. In addition, we also compare the prediction results with emotional attention in Fig. 9, where the weakly supervised model can also match human emotion prioritization.

We also show more detection results on other affective datasets in Fig. 10. As mentioned before, for images with clear foreground and background, some predicted sentiment maps may be similar with the salient regions. However, there are also significantly different regions. For example, the first "fear" image in Fig. 10 highlights the scary face but not the other one. Meanwhile, the sentiment-related background can also be detected in our sentiment maps. For example, the second "positive" image detects the sea in the background rather than the foreground chair. Moreover, for more complex user-contributed images, the detected sentiment maps achieve comparable results to human annotations.

F. Failure Case Analysis

We show some failure detection cases of our framework in Fig. 11. As can be seen, these failure cases can be categorized into the following situations. In the first situation, the detected

region covers the main part of the sentiment regions, while the regions of no interest are also detected resulting in low precision value. Typical examples are the images shown in the first row of Fig. 11. For the second situation, as shown in the second row, the regions evoking sentiment are not completely localized, which can be caused by the low contrast between the foreground and background. Last but not least, as is typical for weakly-supervised methods, the detected results may be misled by the salient objects in the complex scene as shown in the last row.

In addition, we also categorize each of our sentiment maps into one of the following five cases similar to [91]: (i) correct detection (Recall>50%), (ii) hypothesis completely inside ground-truth, (iii) reversed inclusion, (iv) none of the above, but non-zero overlap, and (v) no overlap. For the EmotionROI dataset, we show the frequency of these five cases for each sentiment class and averaged over all classes in Fig. 12. We have the following observations. Most failure cases have low overlap and none of the samples belong to the fifth case due to the soft proposal form. On average, our method predicts the correct localization for about 60.27% images. About 37.54% images are detected with low overlap, and only a few images are detected excessively or partially.

Intuitively, a promising solution is to provide more prior knowledge for the weakly-supervised learning process, such as low-level appearances, mid-level features, and high-level structure and attributes of the input images, so that the regions with similar textures or semantics can be detected simultaneously. Another solution is to design more advanced pooling operation utilizing rich information from both low-level and high-level information in CNNs to deal with challenging inputs with complex scenes.

V. Conclusions

In this paper, we present a weakly supervised framework for both sentiment detection and classification, which addresses the problem of time- and labor-consuming annotation process in this domain. We develop an end-to-end coupled network to take multiple information into consideration, which learns the robust representation with two branches. The detection branch is designed to automatically exploit the sentiment map, which can provide the localized information of the affective images. Then the classification branch, leveraging both holistic and localized representations, can predict the sentiments. Experimental results show the effectiveness of our method against state-of-the-art algorithms on seven benchmark datasets. In addition, analyses on EmotionROI and EMOd show the effectiveness of the weakly supervised sentiment detection.

Our weakly supervised coupled network combining localized information with global representation have applications beyond the proposed method. For example, tasks like the aesthetic evaluation and painting classification are also subjective to obtain sufficient data, while our framework leveraging weakly-supervised localization information can be beneficial for analyzing such subjective tasks. In addition, the WSCNet can also be adapted for semantic applications like Visual Question Answering where localized features are important.

Acknowledgement

This work was supported by the NSFC (No. 61876094, U1933114), Natural Science Foundation of Tianjin, China (No. 18JCYBJC15400, 18ZXZNGX00110), the Open Project Program of the National Laboratory of Pattern Recognition (NLPR), and the Fundamental Research Funds for the Central Universities.

References

[1] J. Yang, D. She, Y.-K. Lai, P. L. Rosin, and M.-H. Yang, "Weakly supervised coupled networks for visual sentiment analysis," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 1, 2

[2] D. Joshi, R. Datta, E. Fedorovskaya, Q.-T. Luong, J. Z. Wang, J. Li, and J. Luo, "Aesthetics and emotions in images," *IEEE Signal Processing Magazine*, vol. 28, no. 5, pp. 94–115, 2011. 1, 2

[3] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Found. Trends Inf. Ret.*, vol. 2, no. 1–2, pp. 1–135, 2008. 1

[4] Q. Truong and H. W. Lauw, "Visual sentiment analysis for review images with item-oriented and user-oriented CNN," in *ACM Int. Conf. Multimedia*, 2017. 1

[5] W. Wang, Y. Yu, and S. Jiang, "Image retrieval by emotional semantics: A study of emotional space and feature extraction," in *Int. Conf. Syst. Man. Cy.*, 2006. 1

[6] J. Yang, D. She, Y.-K. Lai, and M.-H. Yang, "Retrieving and classifying affective images via deep metric learning," in *AAAI Conf. Artif. Intell.*, 2018. 1

[7] J. Jia, S. Wu, X. Wang, P. Hu, L. Cai, and J. Tang, "Can we understand van gogh's mood? learning to infer affects from images in social networks," in *ACM Int. Conf. Multimedia*, 2012. 1

[8] Y.-Y. Chen, T. Chen, T. Liu, H.-Y. M. Liao, and S.-F. Chang, "Assistive image comment robot – A novel mid-level concept-based representation," *IEEE Trans. Affect. Comput.*, vol. 6, no. 3, pp. 298–311, 2015. 1

[9] K.-C. Peng, T. Chen, A. Sadovnik, and A. C. Gallagher, "A mixed bag of emotions: Model, predict, and transfer emotion distributions," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2015. 1

[10] Q. You, J. Luo, H. Jin, and J. Yang, "Building a large scale dataset for image emotion recognition: The fine print and the benchmark," in *AAAI Conf. Artif. Intell.*, 2016. 1, 2, 6

[11] V. Yanulevskaya, J. Van Gemert, K. Roth, A.-K. Herbold, N. Sebe, and J.-M. Geusebroek, "Emotional valence categorization using holistic image features," in *IEEE Int. Conf. Image Process.*, 2008. 1, 2

[12] J. Machajdik and A. Hanbury, "Affective image classification using features inspired by psychology and art theory," in *ACM Int. Conf. Multimedia*, 2010. 1, 2

[13] S. Zhao, Y. Gao, X. Jiang, H. Yao, T.-S. Chua, and X. Sun, "Exploring principles-of-art features for image emotion recognition," in *ACM Int. Conf. Multimedia*, 2014. 1, 2, 7

[14] K.-C. Peng, A. Sadovnik, A. Gallagher, and T. Chen, "Where do emotions come from? predicting the emotion stimuli map," in *IEEE Int. Conf. Image Process.*, 2016. 1, 2, 3, 4, 6, 7, 8, 10

[15] S. Fan, Z. Shen, M. Jiang, B. L. Koenig, J. Xu, M. S. Kankanhalli, and Q. Zhao, "Emotional attention: A study of image sentiment and visual attention," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 1, 2, 3, 6, 8, 10

[16] R. J. Compton, "The interface between emotion and attention: a review of evidence from psychology and neuroscience," *Behav. Cogn. Neurosci. Rev.*, vol. 2, no. 2, pp. 115–129, 2003. 1, 2

[17] R. Gupta, "Commentary: Neural control of vascular reactions: Impact of emotion and attention," *Frontiers in Psychology*, vol. 7, 2016. 1

[18] A. Öhman, A. Flykt, and F. Esteves, "Emotion drives attention: detecting the snake in the grass." *Journal of experimental psychology: general*, vol. 130, no. 3, p. 466, 2001. 1

[19] Q. You, J. Luo, H. Jin, and J. Yang, "Robust image sentiment analysis using progressively trained and domain transferred deep networks," in *AAAI Conf. Artif. Intell.*, 2015. 1, 2, 6

[20] V. Campos, B. Jou, and X. Giró i Nieto, "From pixels to sentiment: Fine-tuning CNNs for visual sentiment prediction," *Image Vision Comput.*, vol. 65, pp. 15–22, 2017. 1, 2

[21] R. Kosti, J. M. Alvarez, A. Recasens, and À. Lapedriza, "Emotion recognition in context," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017. 1

[22] H. Zhang and M. Xu, "Modeling temporal information using discrete fourier transform for recognizing emotions in user-generated videos," in *IEEE Int. Conf. Image Process.*, 2016. 2

[23] J. Gao, Y. Fu, Y.-G. Jiang, and X. Xue, "Frame-transformer emotion classification network," in *ACM Int. Conf. Multimedia Retri.*, 2017. 2

[24] S. Zhao, G. Ding, Y. Gao, and J. Han, "Approximating discrete probability distribution of image emotions by multi-modal features fusion," in *Int. J. Conf. Artif. Intell.*, 2017. 2

[25] S. Zhao, H. Yao, Y. Gao, R. Ji, and G. Ding, "Continuous probability distribution prediction of image emotions via multi-task shared sparse regression," *IEEE Trans. Multimedia*, vol. 19, no. 3, pp. 632–645, 2017. 2

[26] K. Boehner, R. de Paula, P. Dourish, and P. Sengers, "How emotion is made and measured," *Int. J. Hum.-Comput. Stud.*, vol. 65, no. 4, pp. 275–291, 2007. 2

[27] X. Yao, D. She, S. Zhao, J. Liang, Y.-K. Lai, and J. Yang, "Attention-aware polarity sensitive embedding for affective image retrieval," in *Int. Conf. Comput. Vis.*, 2019. 2

[28] C. Zhan, D. She, S. Zhao, M.-M. Cheng, and J. Yang, "Zero-shot emotion recognition via affective structural embedding," in *Int. Conf. Comput. Vis.*, 2019. 2

[29] J. Yuan, S. Mcdonough, Q. You, and J. Luo, "Sentribute: Image sentiment analysis from a mid-level perspective," in *WISDOM*, 2013. 2

[30] A. Sartori, D. Culibrk, Y. Yan, and N. Sebe, "Who's afraid of Itten: Using the art theory of color combination to analyze emotions in abstract paintings," in *ACM Int. Conf. Multimedia*, 2015. 2

[31] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang, "Large-scale visual sentiment ontology and detectors using adjective noun pairs," in *ACM Int. Conf. Multimedia*, 2013. 2, 6, 7

[32] S. Zhao, H. Yao, Y. Gao, R. Ji, W. Xie, X. Jiang, and T. Chua, "Predicting personalized emotion perceptions of social images," in *ACM Int. Conf. Multimedia*, 2016. 2

[33] S. Zhao, H. Yao, Y. Gao, G. Ding, and T. S. Chua, "Predicting personalized image emotion perceptions in social networks," *IEEE Transactions on Affective Computing*, 2018. 2

[34] R. Ji, D. Cao, Y. Zhou, and F. Chen, "Survey of visual sentiment prediction for social media analysis," *Frontiers Comput. Sci.*, vol. 10, no. 4, pp. 602–611, 2016. 2

[35] X. Zhu, L. Li, W. Zhang, T. Rao, M. Xu, Q. Huang, and D. Xu, "Dependency exploitation: A unified CNN-RNN approach for visual emotion recognition," in *Int. J. Conf. Artif. Intell.*, 2017. 2

[36] J. Yang, D. She, and M. Sun, "Joint image emotion classification and distribution learning via deep convolutional neural network," in *Int. J. Conf. Artif. Intell.*, 2017. 2, 5, 7, 8

[37] D. She, M. Sun, and J. Yang, "Learning discriminative sentiment representation from strongly- and weakly-supervised cnns," in *ACM Trans. Multim. Comput.*, 2019. 2

[38] T. Chen, D. Borth, T. Darrell, and S. F. Chang, "Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks," in *arXiv preprint arXiv:1410.8586*, 2014. 2, 7

[39] V. Campos, A. Salvador, X. Giro-i Nieto, and B. Jou, "Diving deep into sentiment: Understanding fine-tuned CNNs for visual sentiment prediction," in *ACM ASM*, 2015. 2

[40] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2009. 2

[41] L. Wu, S. Liu, M. Jian, J. Luo, X. Zhang, and M. Qi, "Reducing noisy labels in weakly labeled data for visual sentiment analysis," in *IEEE Int. Conf. Image Process.*, 2017. 2

[42] T. Rao, X. Li, H. Zhang, and M. Xu, "Multi-level region-based convolutional neural network for image emotion classification," *Neural. Comput.*, vol. 333, pp. 429–439, 2019. 2

[43] M. Katsurai and S. Satoh, "Image sentiment analysis using latent correlations among visual, textual, and sentiment views," in *IEEE Int. Conf. Acou. Speech Signal Process.*, 2016. 2, 6, 7

[44] J. Wu, Y. Yu, C. Huang, and K. Yu, "Deep multiple instance learning for image classification and auto-annotation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2015. 3

[45] B. Zhou, A. Khosla, À. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016. 3, 4, 5, 6, 7, 8, 11

[46] P. Vuilleumier, "How brains beware: Neural mechanisms of emotional attention," *Trends Cong. SCI.*, vol. 9, no. 12, pp. 585–594, 2005. 2

[47] T. Brosch, G. Pourtois, and D. Sander, "The perception and categorisation of emotional stimuli: A review," *Cog. Emotion*, vol. 24, no. 3, pp. 377–400, 2010. 2

[48] B. Li, W. Xiong, W. Hu, and X. Ding, "Context-aware affective images classification based on bilayer sparse representation," in *ACM Int. Conf. Multimedia*, 2012. 3

[49] M. Sun, J. Yang, K. Wang, and H. Shen, "Discovering affective regions in deep convolutional neural networks for visual sentiment prediction," in *Int. Conf. Multimedia and Expo*, 2016. 3, 7, 8

[50] J. Yang, D. She, M. Sun, M.-M. Cheng, P. L. Rosin, and L. Wang, "Visual sentiment prediction based on automatic discovery of affective regions," *IEEE Trans. Multimedia*, 2018. 3

[51] Q. You, H. Jin, and J. Luo, "Visual sentiment analysis by attending on local image regions," in *AAAI Conf. Artif. Intell.*, 2017. 3

[52] S. Fan, M. Jiang, Z. Shen, B. L. Koenig, M. S. Kankanhalli, and Q. Zhao, "The role of visual attention in sentiment prediction," in *ACM Int. Conf. Multimedia*, 2017. 3

[53] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Adv. Neural Inform. Process. Syst.*, 2012. 3, 7

[54] Y. Wei, H. Xiao, H. Shi, Z. Jie, J. Feng, and T. S. Huang, "Revisiting dilated convolution: A simple approach for weakly- and semi- supervised semantic segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 3

[55] X. Zhang, Y. Wei, G. Kang, Y. Yang, and T. Huang, "Self-produced guidance for weakly-supervised object localization," in *Eur. Conf. Comput. Vis.*, 2018. 3

[56] H. Bilen and A. Vedaldi, "Weakly supervised deep detection networks," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016. 3

[57] R. G. Cinbis, J. J. Verbeek, and C. Schmid, "Multi-fold MIL training for weakly supervised object localization," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2014. 3

[58] W. Ren, K. Huang, D. Tao, and T. Tan, "Weakly supervised large scale object localization with multiple instance learning and bag splitting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 405–416, 2016. 3

[59] D. Zhang, D. Meng, and J. Han, "Co-saliency detection via a self-paced multiple-instance learning framework," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 5, pp. 865–878, 2016. 3

[60] R. G. Cinbis, J. Verbeek, and C. Schmid, "Weakly supervised object localization with multi-fold multiple instance learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 1, pp. 189–203, 2017. 3

[61] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Is object localization for free?-weakly-supervised learning with convolutional neural networks," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2015. 3, 5

[62] Y. Zhu, Y. Zhou, Q. Ye, Q. Qiu, and J. Jiao, "Soft proposal networks for weakly supervised object localization," in *Int. Conf. Comput. Vis.*, 2017. 3, 5, 7

[63] P. Dabkowski and Y. Gal, "Real time image saliency for black box classifiers," in *Adv. Neural Inform. Process. Syst.*, 2017. 3

[64] L. Porzi, S. R. Bulò, B. Lepri, and E. Ricci, "Predicting and understanding urban perception with convolutional neural networks," in *ACM Int. Conf. Multimedia*, 2015. 3

[65] X. Alameda-Pineda, A. Pilzer, D. Xu, N. Sebe, and E. Ricci, "Viraliency: Pooling local virality," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 484–492. 3

[66] T. Durand, T. Mordan, N. Thome, and M. Cord, "Wildcat: Weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017. 3, 4, 7, 8, 9

[67] Y. Wei, J. Feng, X. Liang, M.-M. Cheng, Y. Zhao, and S. Yan, "Object region mining with adversarial erasing: A simple classification to semantic segmentation approach," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017. 3

[68] X. Zhang, Y. Wei, J. Feng, Y. Yang, and T. S. Huang, "Adversarial complementary learning for weakly supervised object localization," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 3

[69] H. A. Murray and C. D. Morgan, "A clinical study of sentiments (i & ii)." *Genetic Psychology Monographs*, 1945. 4

[70] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Adv. Neural Inform. Process. Syst.*, 2015. 4, 8

[71] A. Diba, V. Sharma, A. M. Pazandeh, H. Pirsiavash, and L. V. Gool, "Weakly supervised cascaded convolutional networks," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017. 4

[72] D. Zhang, J. Han, L. Chao, J. Wang, and X. Li, "Detection of co-salient objects by looking deep and wide," *Int. J. Comput. Vis.*, vol. 120, no. 2, pp. 215–232, 2016. 4

[73] C. Gong, P. Zhou, and J. Han, "Learning rotation-invariant convolutional neural networks for object detection in vhr optical remote sensing images," *IEEE Trans. Geosci. & Remote Sensing*, vol. 54, no. 12, pp. 7405–7415, 2016. 4
[74] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Eur. Conf. Comput. Vis.*, 2014. 5
[75] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Int. Conf. Learn. Represent.*, 2015. 7
[76] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016. 7
[77] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, and Z. Zhang, "The application of two-level attention models in deep convolutional neural network for fine-grained image classification," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2015. 6
[78] L. Wu, Y. Wang, X. Li, and J. Gao, "Deep attention-based spatially recursive networks for fine-grained visual recognition," *IEEE Trans. on Cyber.*, 2018. 6
[79] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization." in *Int. Conf. Comput. Vis.*, 2017. 6
[80] A. Paszke, S. Gross, S. Chintala *et al.*, "Pytorch," 2017. 7
[81] B. Alexe, T. Deselaers, and V. Ferrari, "Measuring the objectness of image windows," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2189–2202, 2012. 8, 10
[82] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Adv. Neural Inform. Process. Syst.*, 2007. 8
[83] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, 1998. 8
[84] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, p. 27, 2011. 7
[85] X. Huang, C. Shen, X. Boix, and Q. Zhao, "Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks," in *Int. Conf. Comput. Vis.*, 2015. 8
[86] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, "What do different evaluation metrics tell us about saliency models?" *arXiv preprint arXiv:1604.03605*, 2016. 8
[87] O. Le Meur, P. Le Callet, and D. Barba, "Predicting visual fixations on video based on low-level visual features," *Vis. Res.*, vol. 47, no. 19, pp. 2483–2498, 2007. 8
[88] M. J. Swain and D. H. Ballard, "Color indexing," *Int. J. Comput. Vis.*, vol. 7, no. 1, pp. 11–32, 1991. 8
[89] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *Int. J. Comput. Vis.*, vol. 40, no. 2, pp. 99–121, 2000. 8
[90] J. M. Joyce, "Kullback-leibler divergence," in *Int. J. Electrochem. Sc.* Springer, 2011, pp. 720–722. 8
[91] R. G. Cinbis, J. J. Verbeek, and C. Schmid, "Weakly supervised object localization with multi-fold multiple instance learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 1, pp. 189–203, 2017. 12